根据提供的两份录音文件，这位老师（或两位老师）针对机器学习期末考试进行了非常详细的划重点。以下是为您总结的必考点、必考名词解释以及需要掌握的核心内容：

一、必考的名词解释 (Must-Memorize Definitions)

这些概念被明确提到可能会以“名词解释”或填空的形式出现，必须准确记忆定义。

- **泛化能力 (Generalization Ability):** 机器学习算法对新样本的适应能力[1]。
- **支持向量 (Support Vector):** 决策边界上的样本点（也就是距离超平面最近的样本），决定了SVM的划分结果[2]。
- **强化学习 (Reinforcement Learning):** 核心定义是“有延迟标记信息的监督学习”或“试错学习”，要记住这句话[3]。
- **归纳偏好 (Inductive Bias):** 机器学习算法在学习过程中对某种假设的偏好（例如奥卡姆剃刀原则：若非必要，勿增实体/选简单的模型）[4]。
- **错误率与误差 (Error Rate & Error):** 错误样本占比即错误率；真实输出与预测输出的差异即误差（分训练误差、测试误差、泛化误差）[5]。
- **过拟合 (Overfitting):** 把训练样本自身的特点当作所有潜在样本的一般性质（例如认为树叶必须有锯齿），导致泛化能力下降[6]。

二、必考公式与算法推导 (Formulas & Algorithms)

老师明确提到会考“写公式”、“写一般形式”或“写算法流程”的内容。

1. 线性模型 **(Linear Models):**
   - **写公式:** 必须能写出线性模型的一般形式 $f(x) = w^T x + b$[7777]。
   - **推导:** 掌握最小二乘法 **(Least Squares)** 的参数估计推导，特别是求偏导数的过程（老师提到这是最简单的公式，必考）[8]。
   - **简答:** 线性模型的优点[9999]。
2. 决策树 **(Decision Trees):**
   - **写公式:** 信息熵 **(Information Entropy)** 的公式 $Ent(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$。老师强调要能默写出来，并且最好加上对符号的解释（如 p 是概率，求和是对类别求和）[10]。
   - * **理解:** 信息增益 (Information Gain) 的概念，以及信息增益率 (Gain Ratio) 是为了

解决什么问题（惩罚取值过多的属性）[cite: 1614]。

3. 朴素贝叶斯 **(Naïve Bayes):**
    - 写公式**:** 写出朴素贝叶斯公式 $P(c|x) \propto P(c) \prod_{i=1}^d P(x_i|c)$ 并解释其中 $P(c)$ 是先验概率，$P(x|c)$ 是条件概率/似然 [11]。
    - 概念**:** 拉普拉斯平滑 (Laplacian Smoothing) 的作用（避免零概率）[12]。
4. **AdaBoost** 算法**:**
    - 老师在录音中专门讨论了考 AdaBoost 还是考 K-Means，最终倾向于考 AdaBoost 的算法流程或核心思想（串行生成，关注错分样本）[13131313]。

三、重点必考板块 (Key Topics for Choice/Short Answer)

这些内容主要出现在选择题、简答题或综合题中。

1. 机器学习基础与评估

- 发展阶段**:** 必须搞清楚人工智能发展的三个阶段顺序（推理期 -> 知识期 -> 学习期），以及三个流派（符号主义 -> 连接主义 -> 统计学习 -> 连接主义复兴/深度学习 ）的代表技术和时间顺序 [14]。
- 评估方法**:**
    - 交叉验证法 **(Cross-Validation):** 必考，要记住 [15]。
    - **P-R** 曲线与 **ROC** 曲线**:** 知道如何判断模型好坏。P-R曲线看平衡点 **(Break-Event Point)** 越靠外越好；ROC曲线看曲线下的面积 (AUC) 越大越好，若曲线被包围则性能较差 [16161616]。
- 偏差-方差分解 **(Bias-Variance Decomposition):** 理解偏差、方差、噪声的含义，以及训练程度（欠拟合/过拟合）与偏差方差的关系 [17]。
- 假设检验**:** 为什么需要？（消除随机性，保证结果可靠）[18]。

2. 支持向量机 (SVM)

- 核心目标**:** 学习线性可分（或近似可分）数据的分类超平面 [19191919]。
- 间隔最大化**:** 知道什么是硬间隔 (Hard Margin) [20]。
- 求解方法**:** 知道引入拉格朗日乘子 **(Lagrange Multipliers)** 和对偶问题 [21]。

3. 聚类与集成学习

- **K-Means:** 算法步骤、缺点以及改进方法 [22]。
- 集成学习**:** 理解 Bagging 和 Boosting 的区别（并行 vs 串行），理解为什么集成学习有效（基学习器要有差异性，“好而不同”）[23232323]。

4. 深度学习与大模型 **(**前沿考点**)**

- 大模型数据处理四阶段**:** 规则去重 -> 模型过滤 -> 隐私安全（记住这几个词 ） [24]。
- 训练流程**:** 预训练 (Pre-training) -> 有监督微调 (SFT) -> 强化学习 (RLHF) [25]。
- 变形**/**加速框架**:** 知道数据并行、流水线并行、张量并行等概念 [26]。
- **Scaling Law:** 模型做大做强的好处（跳出局部极小值，方差变小） [27]。

5. 强化学习 **(RL)**

- 元组**:** 解释 $<S, A, P, R>$ 分别代表什么（状态空间、动作空间、转移概率、奖赏 ） [28]。
- 困境**:** 探索 (Exploration) 与利用 (Exploitation) 的平衡 [29]。

四、老师给出的特别复习建议

1. 不要留白**:** 即使不会写，也不要交白卷，写上相关公式或概念可能会给分 [30]。
2. **PPT** 重点**:** 老师提到重点 PPT 大概有 20-40 页，灰色标记的不会考，黄色标记的必考（通常是简答或大题 ），白色背景的用于辅助理解 [3131313131313131]。
3. 记忆优先**:** 期末考试被老师定义为“记忆的游戏”，对于无法完全理解的算法细节（如复杂的推导 ），建议先把结论和关键步骤背下来 [32323232]。

下一步建议：您可以开始针对上述“必考公式”进行默写练习，尤其是线性模型的最小二乘法推导和信息熵公式。需要我为您详细展开某个特定公式的推导过程吗？